Predicting Growth Traits with Genomic Selection Methods in Zhikong Scallop (Chlamys farreri)

Wang, Yangfan; Sun, Guidong; Zeng, Qifan; Chen, Zhihui; Hu, Xiaoli; Li, Hengde

Published in:
Marine Biotechnology

DOI:
10.1007/s10126-018-9847-z

Publication date:
2018

Document Version
Peer reviewed version

Predicting Growth Traits with Genomic Selection Methods in Zhikong Scallop (*Chlamys farreri*)

Yangfan Wang, [1]

Guidong Sun, [1]

Qifan Zeng, [1, 2]✉

Email zengqifan@ouc.edu.cn

Zhihui Chen, [~~2~~3]

Xiaoli Hu, [1,~~3~~4]

Hengde Li, [~~4~~5]

Shi Wang, [1,~~5~~2]

Zhenmin Bao, [1,~~3~~4]

[1] Ministry of Education Key Laboratory of Marine Genetics and Breeding, College of Marine Science, Ocean University of China, Qingdao, 266003 China AQ1

[~~2~~3] Division of Cell and Developmental Biology, College of Life Science, University of Dundee, Dundee, DD1 4HN UK

[~~3~~4] Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao National Laboratory for Marine Science and Technology, Qingdao, 266237 China

[~~4~~5] Ministry of Agriculture Key Laboratory of Aquatic Genomics, CAFS Key Laboratory of Aquatic Genomics and Beijing Key Laboratory of Fishery Biotechnology, Center for Applied Aquatic Genomics, Chinese Academy of Fishery Sciences, Beijing, 100141 China

[~~5~~2] Laboratory for Marine Biology and Biotechnology, Qingdao National Laboratory for Marine Science and Technology, Qingdao, 266237 China

This is a post-peer-review, pre-copyedit version of an article Wang, Yangfan et al."Predicting Growth Traits with Genomic Selection Methods in Zhikong Scallop (Chlamys farreri)". Marine Biotechnology. 2018. Available: 10.1007/s10126-018-9847-z

Received: 6 June 2018 / Accepted: 29 July 2018

Abstract

Selective breeding is a common and effective approach for genetic improvement of aquaculture stocks with parental selection as the key factor. Genomic selection (GS) has been proposed as a promising tool to facilitate selective breeding. Here, we evaluated the predictability of four GS methods in Zhikong scallop (*Chlamys farreri*) through real dataset analyses of four economical traits (e.g., shell length, shell height, shell width, and whole weight). Our analysis revealed that different GS models exhibited variable performance in prediction accuracy depending on genetic and statistical factors, but non-parametric method, including reproducing kernel Hilbert spaces regression (RKHS) and sparse neural networks (SNN), generally outperformed parametric linear method, such as genomic best linear unbiased prediction (GBLUP) and BayesB. Furthermore, we demonstrated that the predictability relied mainly on the heritability regardless of GS methods. The size of training population and marker density also had considerable effects on the predictive performance. In practice, increasing the training population size could better improve the genomic prediction than raising the marker density. This study is the first to apply non-linear model and neural networks for GS in scallop and should be valuable to help develop strategies for aquaculture breeding programs.

Keywords

Genomic selection
Heritability
Breeding
Scallop

Introduction

Selective breeding is a common and effective approach for genetic improvement through choosing parents with desired characteristics. Traditional aquaculture selection methods, such as sib-testing, have limited reliability due to that selection candidates are evaluated based on mid-parent means (Odegard et al. 2014). Furthermore, classical selection schemes also lead to increased co-selection among close relatives and applying constraints on inbreeding hinder selection on the interested traits rather than selection for individually evaluated traits (Rodríguez-Ramilo et al. 2015).

Genomic selection has been proposed as a promising tool to facilitate selective breeding (Meuwissen et al. 2001). The basic concept of GS is to estimate the marker effect in a training population and to predict the genomic estimated breeding value (GEBV) of selection candidates. Compared with traditional marker assisted selection method, GS requires no significant test, therefore, avoids biases in marker effect estimates and could accelerate the breeding cycle (Goddard and Hayes 2009; Hill 2013). Because of its high prediction accuracy, GS has been widely used in agricultural plants (e.g., Bernardo and Yu 2007; Piepho 2009; Jannink et al. 2010; Crossa et al. 2014) and animals (Gonzalez-Recio et al. 2008; VanRaden 2008; Hayes et al. 2009; de los Campos et al. 2009a). The rapid generation of extensive genomic resources also enabled genomic dissection of complex traits and application of GS in aquaculture species (Ge et al. 2015; Kessuwan et al. 2016; Liu et al. 2016; Abdelrahman et al. 2017; Negrín-Báez et al. 2016; Lin et al. 2018; Sawayama et al. 2017, 2018; Wang et al. 2017a, b, c; Zhong et al. 2017; Li et al. 2018; Zhao et al. 2018). So far, genomic selection has been applied in few aquatic animals, including the large yellow croaker (Dong et al. 2016), the Atlantic salmon (Tsai et al. 2015), the rainbow trout (Vallejo et al. 2016), and the Japanese Flounder (Liu et al. 2018).

AQ2

AQ3

AQ4

AQ5

The prediction performance is essential for successful application of GS. Several factors affecting the prediction performance such as genetic trait architecture, span of linkage disequilibrium (LD), sample size, trait heritability, and marker density have been identified (Zhong et al. 2009; Daetwyler et al. 2010; Habier et al. 2007). In general, the predictability increases as marker intensity and sample size increases until reaches a plateau. The required marker density is determined by the speed of linkage disequilibrium (LD) decays in the population. When LD decays slowly, only a small number of markers could represent the genome (Desta and Ortiz 2014). The predictability is also closely related to the heritability. The traits with higher heritability tend to have higher predictability. The predictabilities of low heritability traits, such as yield, were consistently lower than high heritability traits (Goddard and Hayes 2009). In addition to genetic factors, statistical models in GS have influence on the predictability. Currently, commonly used parametric GS methods include genomic best linear unbiased prediction (GBLUP) (Meuwissen et al. 2001), Bayesian methods (Goddard and Hayes 2009), least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996), and partial least squares (PLS) (Gelandi and

Kowalski 1986). These parametric models have defects because they typically ignore complicated gene interactions or higher order non-linearity relationships in determining marker effect. To take possible non-linearity into account in prediction, it has emerged as a new tool for marker-based genomic predictions of complex traits through non-parametric methods including support vector machine (SVM) (Maenhout et al. 2007), reproducing kernel Hilbert spaces regression (RKHS) (de Los Campos et al. 2009b), and neural networks (NN) (Gianola et al. 2011; Wang et al. 2018).

AQ6

So far, selective breeding in scallop has been performed mainly through traditional selection method. Significant genetic gains from selection for growth have been reported in the Catarian scallop (*Argopecten ventricosus*) (Ibarra et al. 1999), the Japanese scallop (*Patinopecten yessoensis*) (Liang et al. 2010), and the Bay scallop (*Argopecten irradians irradians*) (Zheng et al. 2006). With the development of new genotyping technologies and recently generated genome references (Wang et al. 2016, Li et al. ~~2017a,~~ b, Wang et al. 2017c, d~~Wang et al. 2017a, b, c~~), genomic selection becomes applicable for scallops. Despite that the prediction performances of six parametric GS models have been evaluated in Yesso scallop (Dou et al. 2016), the predictability of non-parametric models as well as their dependent genetic and statistical factors are largely unknown. To demonstrate the utility of GS in scallop selective breeding, we evaluated the accuracy of genomic prediction in an admixed population of Zhikong scallop (*Chlamys farreri*) using RKHS (de Los Campos et al. 2009b) and sparse neural networks (SNN) (Wang et al. 2018). Their performances were compared with traditional method include GBLUP and Bayes B under various conditions. We also assessed the influence of heritability, marker density, and training population size on predicting performance.

AQ7

Results

The SNP-Based Heritability Estimation

The basic SNPs (26,471 SNPs with MAF > 2%) were used for heritability estimation using GCTA. The SNP-based heritability h^2_{GCTA} of shell length, shell height, shell width, and whole weight was 0.42 (S.E. 0.09), 0.47 (S.E. 0.07), 0.54 (S.E. 0.11), and 0.28 (S.E. 0.03), respectively (Table 1). h^2_{GCTA} calculated from 20,000 or 10,000 subsampled SNPs were very close to those from the whole set of SNPs (31,361 SNPs) (Fig. 1). h^2_{GCTA} calculated using 2500 SNPs were significantly lower in all traits, suggesting that insufficient markers could reduce the accuracy. To test the effects of causative SNPs on heritability

estimation, we also excluded 1000 SNPs that GWAS identified as mostly closely associated with phenotypic variance for h^2_{GCTA} calculation. The calculated h^2_{GCTA} remains stable and consistent with reduced markers (Table 1). The results reinforced that the SNP-based estimates do not require the information of major loci for heritability estimation, as long as the SNP density is eligible to capture the fine-scale relatedness.

Table 1

Estimates of h^2_{GCTA} for four traits using basic SNPs with MAF > 2%, common SNPs with MAF > 10%, and with the top 1000 major SNPs masked (MAF > 2%). Standard errors for h^2_{GCTA} estimates are in parentheses

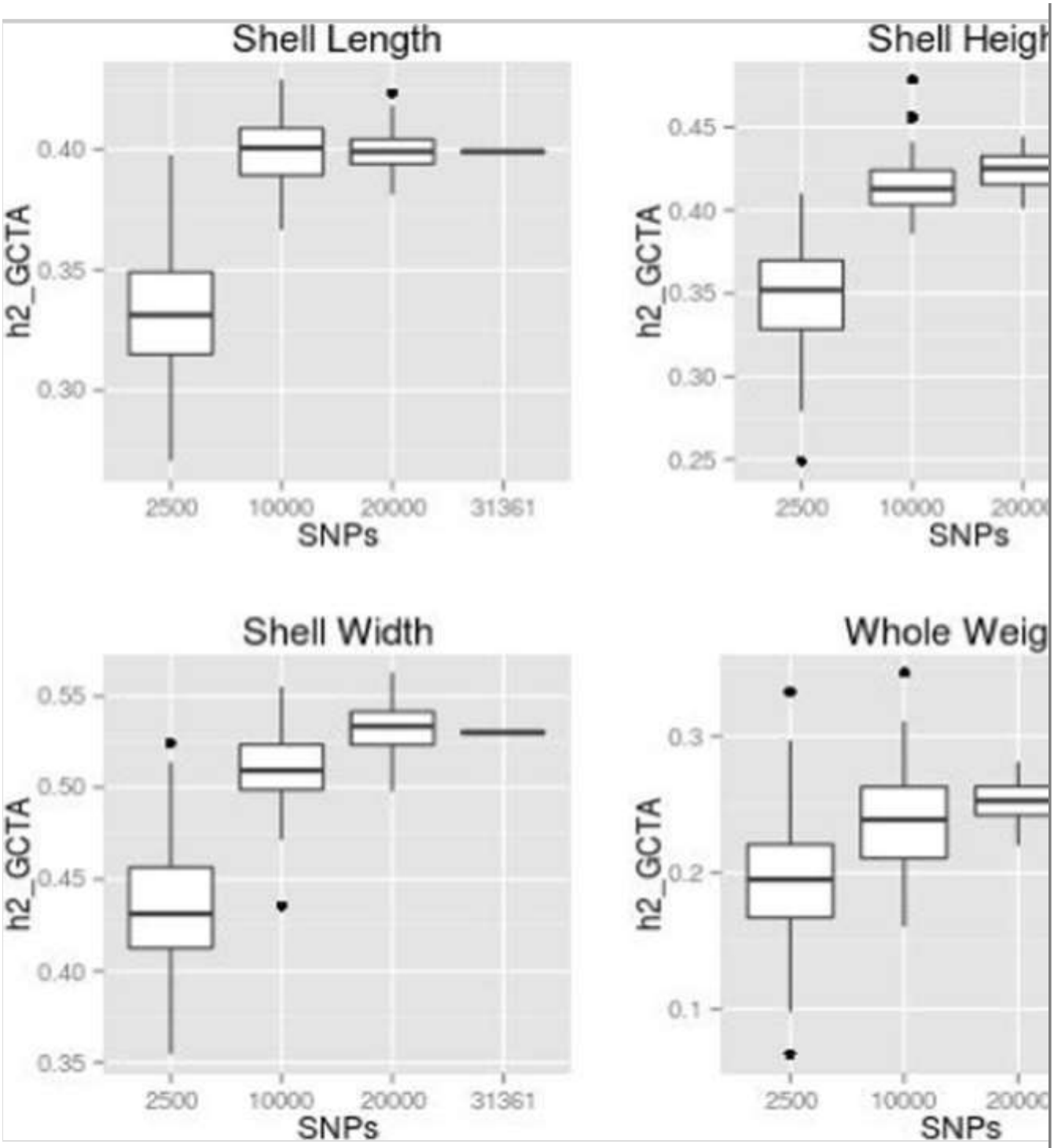
AQ8

Trait	h^2_{GCTA} (S.E.)	h^2_{GCTA} (S.E.)	h^2_{GCTA} (S.E.)
	MAF > 2%	MAF > 10%	Major SNPs masked
Shell length	0.42(0.09)	0.39(0.08)	0.42(0.08)
Shell height	0.47(0.07)	0.43(0.09)	0.46(0.07)
Shell width	0.54(0.11)	0.50(0.12)	0.53(0.11)
Whole weight	0.28(0.03)	0.26(0.05)	0.28(0.03)

Fig. 1

Box-and-whisker plots of SNP-based heritability estimates from 100 samples each made with the 509 scallops at 24 months in the selected groups and 2500, 10,000, 20,000, or 31,361 SNPs for shell length, shell height, shell width, and whole weight

AQ9



Evaluation of the Predictive Power

We have compared the predictive performance of GBLUP, Bayes B, RKHS, and SNN with the scallop data sets. Table 2 presents the evaluation of the predictive performance of the models using basic SNPs with a tenfold cross-validation. The predictabilities of the four models are generally correlated to the trait heritability. As revealed by Table 2, shell width exhibits the highest predictability across all methods, follows by shell height, shell length, and whole weight. Despite the correlation coefficients of whole weight are lower than 0.38, the predictabilities of most traits using different models are all above than 0.42. Based on optimal GS models, the prediction accuracy for this empirical dataset

could reach about 0.51, 0.56, 0.58, and 0.37 for shell length, shell height, shell width, and whole weight, respectively. Different models also exhibited slightly differences for particular traits. The largest differences in predictability among the four methods vary from 0.0333 to 0.1047. Standard deviations of predictabilities range from 0.0076 to 0.0173 across traits and methods, where the high predictable traits tend to have smaller standard deviations than those low predictable traits. Among the four methods, RKHS and SNN generally outperformed GBLUP and Bayes B. For traits including hell height, shell length, and shell width, RKHS is the most efficient method. While for whole weight, SNN is the most efficient instead.

Table 2

Correlations between observed and predicted values for scallop dataset for four traits with different SNP-based heritabilities

Fold	GBLUP	Bayes B	RKHS	SNN	GBLUP	Bayes B	RKHS	SNN
	Shell ~~height~~length with h^2_{GCTA} = 0.42				Shell width with h^2_{GCTA} = 0.54			
1	0.4064	0.4316	0.5056	0.5098	0.4774	0.5064	0.5805	0.5815

COR correlation

Fold	GBLUP	Bayes B	RKHS	SNN	GBLUP	Bayes B	RKHS	SNN
2	0.4241	0.4487	0.5283	0.5289	0.4681	0.5173	0.5847	0.5793
3	0.4377	0.4521	0.5185	0.5069	0.4906	0.5092	0.5656	0.5785
4	0.4427	0.4667	0.5358	0.5405	0.4891	0.5063	0.5956	0.5773
5	0.4342	0.4580	0.5012	0.5114	0.4892	0.5291	0.5741	0.5792
6	0.4275	0.4519	0.5170	0.5138	0.4729	0.5036	0.5745	0.5879
7	0.4132	0.4386	0.5152	0.5088	0.4715	0.5119	0.5668	0.5954
8	0.4078	0.4344	0.5104	0.5071	0.4948	0.5086	0.5785	0.5939
9	0.4181	0.4331	0.5296	0.5118	0.4782	0.5153	0.5878	0.5968
10	0.4188	0.4438	0.5332	0.5329	0.4781	0.5232	0.5708	0.5868
Avg COR	0.4231	0.4459	0.5195	0.5172	0.4810	0.5131	0.5799	0.5857
Sd COR	0.0125	0.0117	0.0119	0.0122	0.0092	0.082	0.0095	0.0076
	Shell height with h^2_{GCTA} = 0.47				Total weight with h^2_{GCTA} = 0.28			

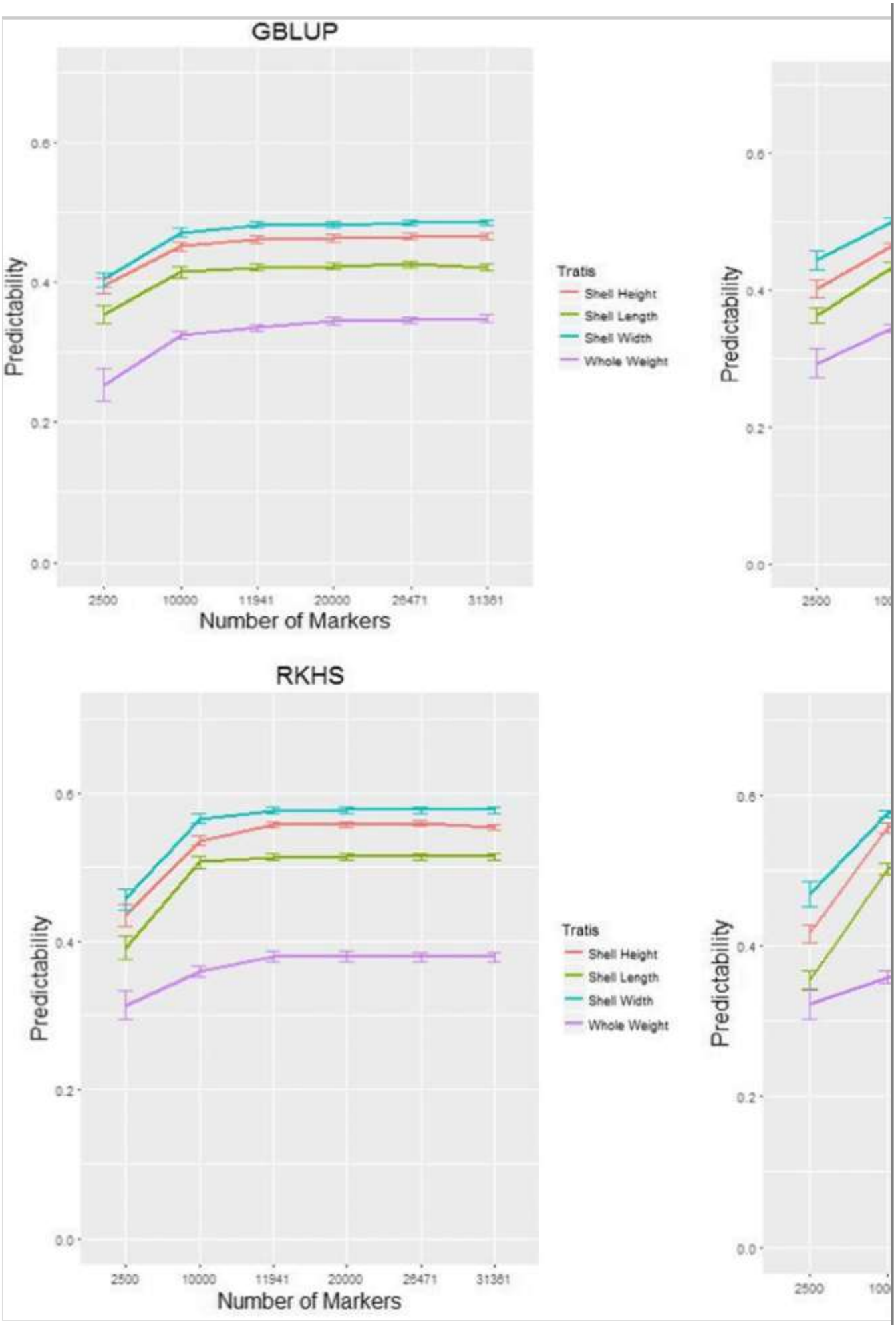
1	0.4657	0.4686	0.5657	0.5648	0.3288	0.3439	0.3802	0.3942
2	0.4733	0.4891	0.5784	0.5679	0.3271	0.3324	0.3851	0.3721
3	0.4743	0.4882	0.5720	0.5474	0.3548	0.3617	0.3710	0.3832
4	0.4504	0.4821	0.5706	0.5786	0.3268	0.3481	0.3445	0.3457
5	0.4482	0.4688	0.5493	0.5663	0.3426	0.3531	0.3866	0.3581
6	0.4589	0.4722	0.5585	0.5498	0.3528	0.3734	0.3834	0.3594
7	0.4714	0.4548	0.5749	0.5591	0.3598	0.3212	0.3573	0.3745
8	0.4611	0.4783	0.5464	0.5857	0.3294	0.3548	0.3704	0.3733
9	0.4448	0.4786	0.5540	0.5564	0.3587	0.3484	0.3955	0.3974
10	0.4671	0.4579	0.5639	0.5704	0.3531	0.3785	0.3934	0.3926
Avg COR	0.4615	0.4739	0.5634	0.5645	0.3434	0.3515	0.3767	0.3751
Sd COR	0.0107	0.0116	0.0111	0.0120	0.1411	0.0173	0.0161	0.0171
COR correlation								

Influence of Marker Number and Training Population Size on Predictability

In order to determine the effect of marker types and densities used for GS in scallop, we selected six subsets of markers including common SNPs (11,941 SNPs with MAF > 10%), basic SNPs (26,471 SNPs with MAF > 2%), randomly sampled SNPs (2500, 10,000, and 20,000 SNPs), and all high-quality SNPs (31,361 SNPs). One hundred selections in a random way were carried out with each subset. As shown in Fig. 2, the predictability remains consistent with over 10,000 SNPs. When the number of markers falls below 10,000, the predictabilities begin to decrease significantly for all traits. The result also reveals that the smaller the number of markers, the larger the variation in predictabilities.

Fig. 2

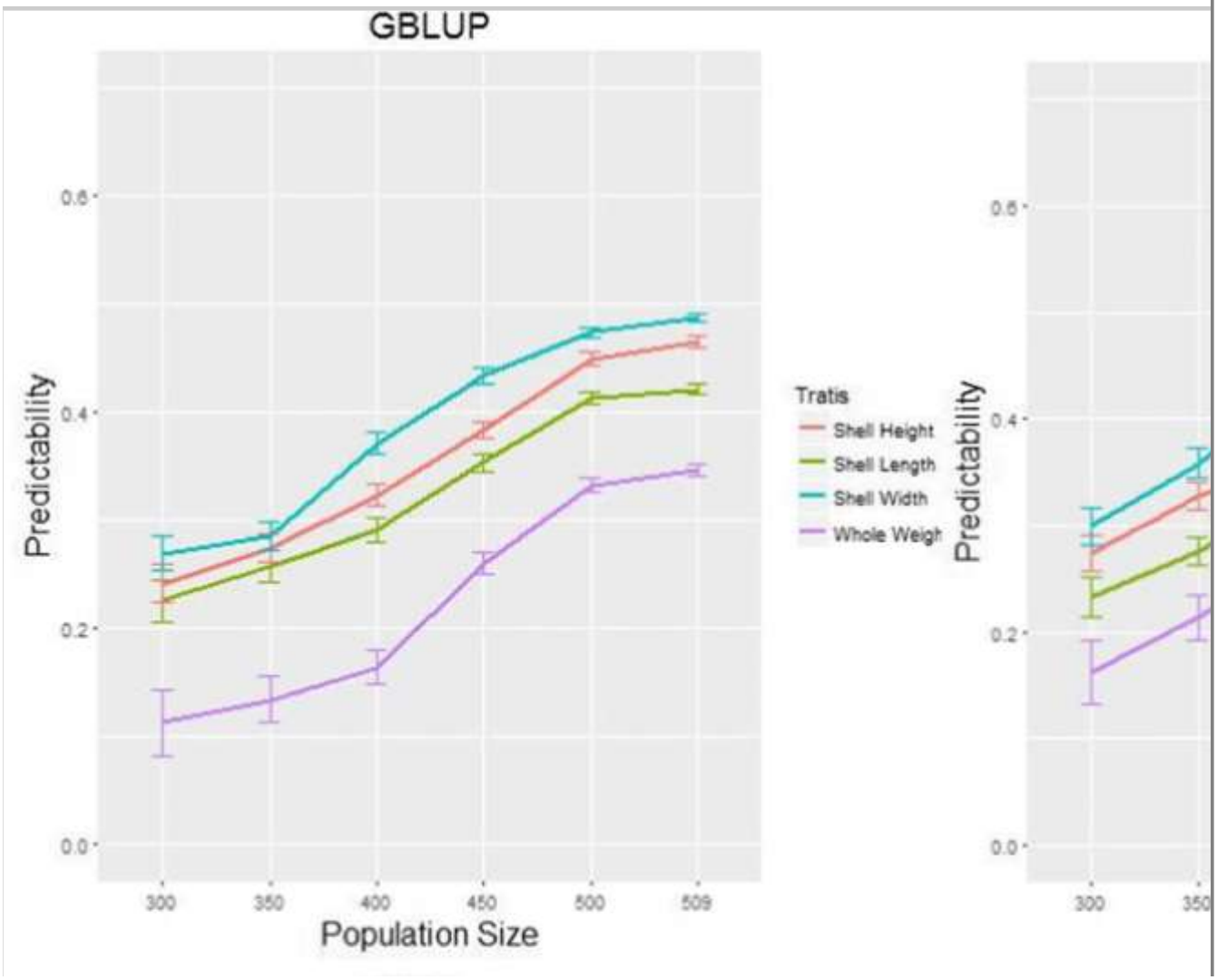
Effect of marker density on the predictability. Six SNP subsets were selected using randomly sampled SNPs (2500, 10,000, and 20,000 SNPs), common SNPs (11,941 SNPs with MAF > 10%), basic SNPs (26,471 SNPs with MAF > 2%), and all high-quality SNPs (31,361 SNPs). Tenfold cross-validations are repeated 100 times for each subset of SNP markers. Error bars are constructed using one standard error from the mean

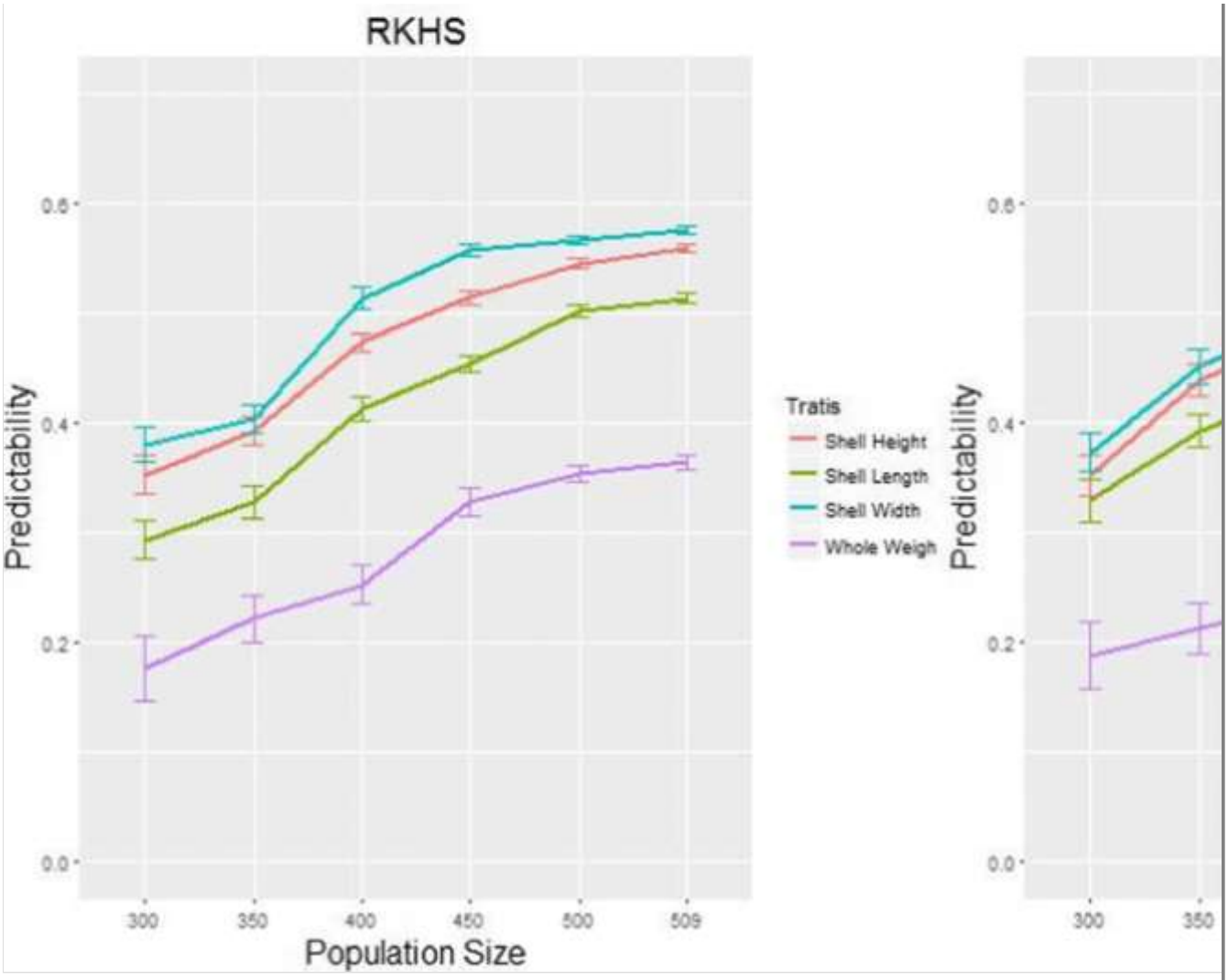

GBLUP
Predictability
0.0
0.2
0.4
0.6
2500
10000
11941
20000
26471
31361
Number of Markers
Tratis
Shell Height
Shell Length
Shell Width
Whole Weight
RKHS
Predictability
0.0
0.2
0.4
0.6
2500
10000
11941
20000
26471
31361
Number of Markers
Tratis
Shell Height
Shell Length
Shell Width
Whole Weight

To investigate the impact of population size on the predictive power of the models, we selected five subsets of training populations varying from 300 individuals to the total 509 individuals. As the size of training population increases from 300 to 509, the average predictabilities of the four methods (GBLUP, Bayes B, RKHS, and SNNR) increased averagely by 47.64%, 43.15%, 41.78%, and 60.23% for the four traits, respectively (Fig. 3). Although both the marker density and the size of training population have influences on the predictability, increasing the training population size could better improve the genomic prediction than raising the marker density. For example, as the number of makers decreases from 31,361 to 2500, the predictabilities of four traits only decline by 11.65% on average, whereas the predictabilities drop by 48.31% on average as the population size decreases from 509 to 300, which indicates that a large training population is necessary to obtain high predictability.

Fig. 3

Effect of the population size on the predictability. Six subsets are selected with the number of individuals varying from 300 to 509 using 10,000 SNP. Tenfold cross-validations are repeated 100 times for each subset of the population. Error bars are constructed using one standard error from the mean





Discussion

In this work, we have evaluated the influences of the GS method, heritability, marker number, and training population size on predictive performance for an admixed population of Zhikong scallop. From the comparison of different prediction methods, we found that non-parametric methods (RKHS, SNN) performed better than parametric methods (GBLUP, BayesB) for the real dataset of scallop. Our results were in consistence with previous studies. Heslot et al. (2012) compared the performance of six parametric methods with four non-parametric methods for genomic prediction in wheat, maize, and barley and observed that the RKHS method performed the best across different species. Ehret et al. (2015) investigated various Bayesian neural network architectures using for predicting phenotypes in Holstein-Friesian and German Fleckvieh cattle and suggested that neural networks can capture non-linearities and may be useful for predicting complex traits using real data. Howard et al. (2014) assessed many parametric and non-parametric methods using simulated genetic architectures, and found that parametric methods performed slightly better than non-parametric methods for additive genetic architectures, but parametric methods had difficulty in capturing non-additive effects such as epistatic effects.

Generally, GBLUP is the most robust method and generally gives the higher predictability for highly polygenic traits; the Bayesian methods are better for traits with major genes; RKHS and SNN perform well for traits under non-additive genetic architectures. If the genetic architecture underlying the trait is unclear, both parametric and non-parametric methods should be tried to cross-confirm the results.

We also found that the size of training population had a greater impact on predicting performance than the marker density did, which was in accordance with earlier studies (Ehret et al. 2015). The increase in predictability quickly reaches a plateau as the number of markers increases. In our study, the predictability plateaued when 10,000 markers were used for prediction of all traits. Research in an elite scallop breeding population genotyped with 2364 markers revealed that prediction accuracy for real dataset of scallop could reach over 0.4 based on optimal GS methods (Dou et al. 2016). The optimal GS methods using 10,000 markers in this work produced the most accurate predictive ability about 18% greater than GS models using only 2364 markers for scallop in Dou et al. 2016. Therefore, a low-density marker panel is desired to obtain a favorable cost-benefit ratio for GS. With respect to the size of training population, it has strong effect on the predictability. We observed a monotonic increase in the predictability for each trait with enhancing population size. Therefore, increasing the size of training population rather than increasing the marker number can be preferable for scallop GS prediction.

Currently, researches on GS are mainly based on the additive model. However, a few studies have suggested that incorporating dominance can produce higher predictability than only considering additive effects (Vitezica et al. 2013). Our result reveals that additive variances may not explain the majority of the trait variances, and the improvement in predictability by including non-additive variances, such as dominance variances, could be considered in near future. This result is consistent with Wang et al. 2018, who found that on average, inclusion of the dominance component yielded better predictions for milk yield supported by results of non-additive effects on milk yield in Jersey cows (Aliloo et al. 2015).

Conclusions

We used an admixed population of Zhikong scallop consisting of 509 individuals to evaluate the genetic and statistical factors affecting prediction in scallop. The results showed that predictabilities for different methods were significantly different, with the two non-linear methods (e.g., RKHS and SNN) better than the two other linear methods. The predictability is closely related to the heritability.

The traits with higher heritability tend to have higher predictability. The size of training population had greater influence on predictive performance compared with the marker density. Our results hold great promise for the implementation of GS in scallop breeding.

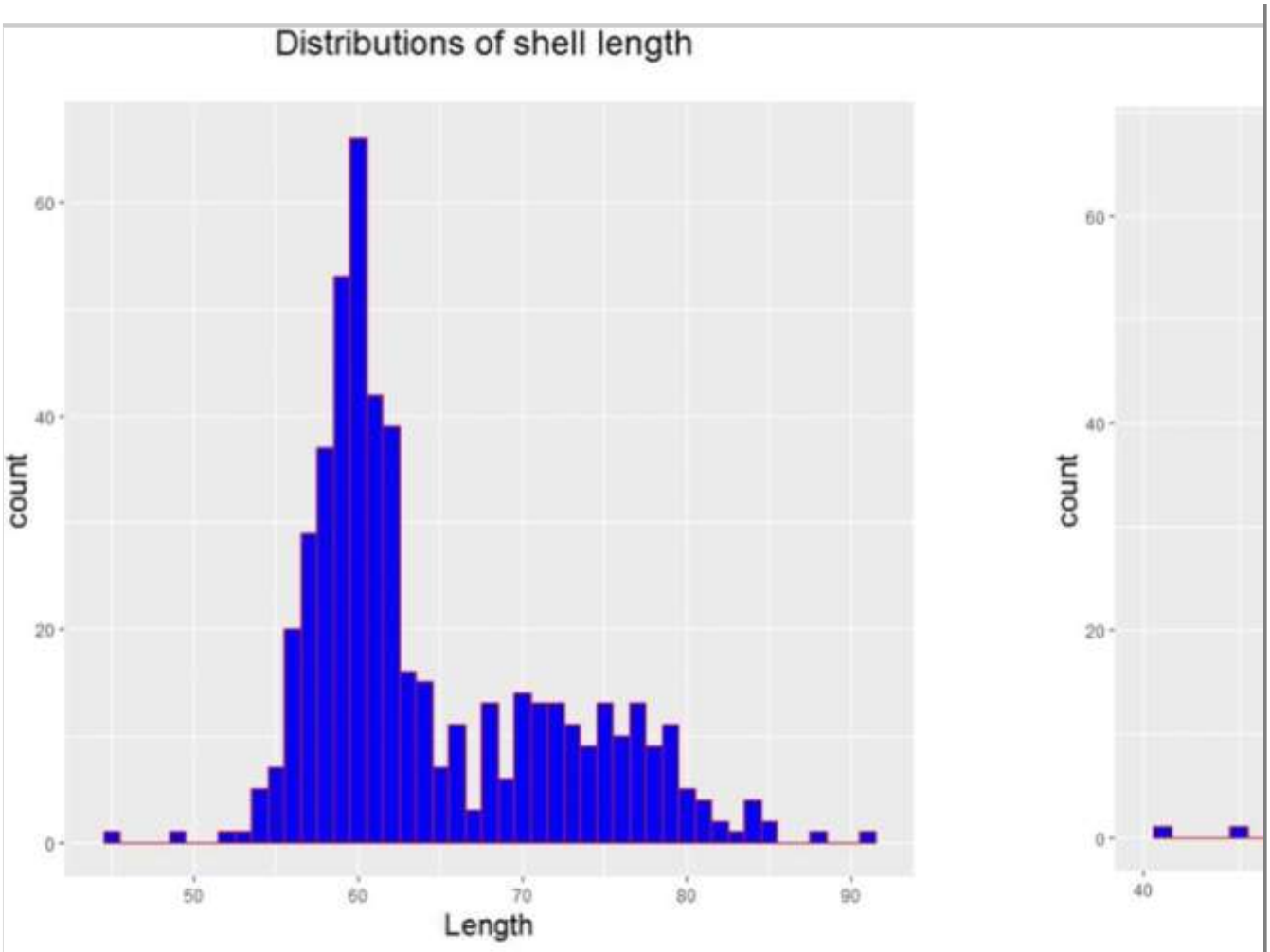
Methods

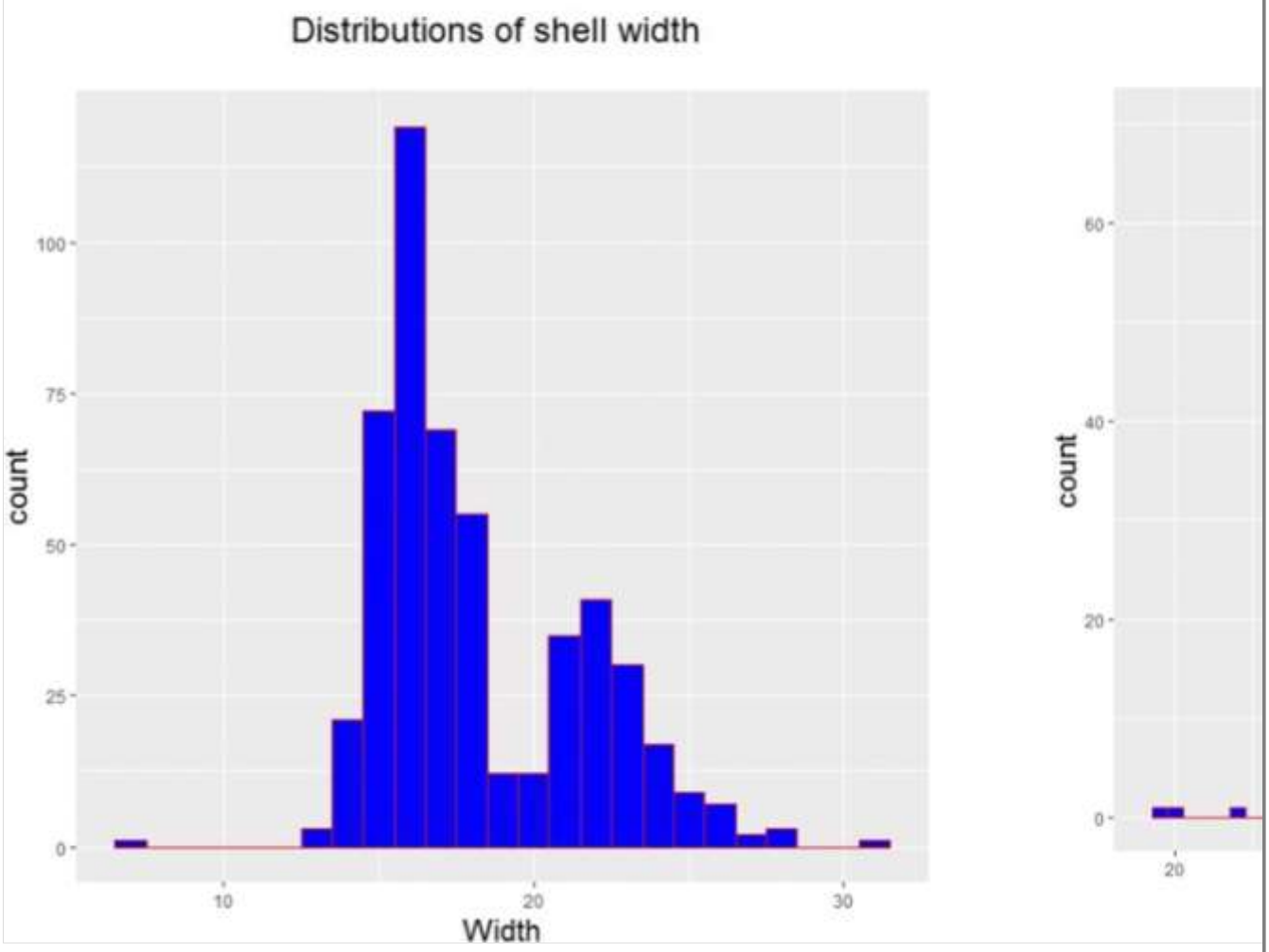
Materials Collection

The Zhikong scallop (*C. farreri*) naturally distributes along the seacoasts of China, Japan, and Korea and is a commercially important bivalve species in China. Currently, genetic studies focusing on scallop growth, reproduction, and immunity represent active research directions. Phenotype data were traditional size-related characters as complex traits, such as shell length, shell height, shell width, and whole wet weight. Shell height was measured from the hinge to the opposite end of the shell. Shell length was measured as the maximum dimension at right angles to the height. Shell width was measured as the greatest vertical distance between the two valves. As shown in Fig. 4, the distributions of shell length, shell height, and shell width were similar and deviated normally distributed. While the distributions of whole weight be approximately normally distributed.

Fig. 4

Distributions of the phenotypes (shell length, shell height, shell width, and whole weight)


Distributions of shell length
count
Length
0
20
40
60
50
60
70
80
90


Distributions of shell width
count
Width
0
25
50
75
100
10
20
30

The parental scallops used in this study were collected from a cultured population in Qingdao Shazikou, Shandong Province, China. In February 2012, 1000 scallops for each trait were brought to the hatchery for selection and conditioning. For the parental populations, the selection intensity planned was $i = 1.755$ for the four complex traits (Falconer and Mackay 1996). However, the observed selection intensity, which was estimated from the standardized difference between the means of the selected parents from the population divided by the standard deviation of the population, was lower for the four traits, 1.651 for shell length, 1.647 for shell height, 1.732 for shell width, and 1.606 for whole wet weight. We randomly collected 509 individuals at 24 months in the selected groups of an admixed population and used 2b-RAD sequencing (Wang et al. 2012) to obtain a high-quality set of SNPs (31,361) with an average calling rate of 84% in this study. Using the physical map of Zhikong scallop (Jiao et al. 2014), the missing genotypes were inferred by the Beagle software (Browning and Browning 2009). Imputation accuracy was measured using R square allelic (R^2), as described by Browning and Browning (2009). We could obtain allelic R^2 for each imputed marker in a sample of 509 individuals. The average and standard error of allelic R^2 values were 0.9084 and 0.1203, respectively.

AQ10

Models of Prediction

GBLUP Meuwissen et al. (2001) introduced the use of linear regression models in genome-enabled predictions. The basic linear regression model for additive effects is

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \qquad 1$$

where y_i is a target trait measured on individual i; μ is an intercept; β_j is the allele substitution effect of marker j and $\beta_j \sim NIID\left(0,\ \sigma^2_{\beta_j}\right)$, where $\sigma^2_{\beta_j}$ is the marker variance; x_{ij} is the jth marker genotype observed in individual i; and $\epsilon_i \sim NIID\left(0,\ \sigma^2_e\right)$, where σ^2_e is the residual variance. GBLUP method assumes $\sigma^2_{\beta_j} = 1/p\sigma^2_\beta$, where σ^2_β is the polygenic variance shared by all makers. The variance-covariance matrix is

$$\mathrm{var}(\mathrm{y}) = \mathbf{V} = \mathbf{W}\mathbf{W}^T / \left(2\sum_{i=1}^{p} p_i\left(1-p_i\right)\right)\ \ \sigma^2_\beta + \mathbf{I}\sigma^2_e = \left(\mathbf{G}\lambda + \mathbf{I}\right)\sigma^2_e \qquad 2$$

where $\lambda = \sigma_\beta^2/\sigma_e^2$ is the signal-noise variance ratio, $\mathbf{W}$ is a standardized genotype matrix with the ij^{th} element $w_{ij} = (x_{ij} - 2p_i)$, p_i is the minor allele frequency for SNP i, and $\mathbf{G}$ is a genomic relationship matrix suggested by VanRaden (2008) and can be written as

$$\mathbf{G} = \mathbf{WW^T} / \left(2\sum_{i=1}^{p} p_i\left(1 - p_i\right)\right). \quad 3$$

BayesB For the BayesB method, we followed Meuwissen et al. (2001). The prior for marker β_j for $j = 1, \ldots, p$ is given by the hierarchical prior

$$\begin{aligned} \beta_j \mid \sigma_{\beta_j}^2 &\sim NIID\left(0,\ \sigma_{\beta_j}^2\right), \\ \sigma_{\beta_j}^2 &\sim \pi\delta_0(.) + (1-\pi)\,\chi^{-2}\left(\upsilon, S\right), \end{aligned} \quad 4$$

where $\delta_0(.)$ denotes a point mass at zero that assigns zero variance to the effects of a fraction π of markers. A priori, only a fraction $1 - \pi$ of markers was selected to be in the model and a scaled inverted chi-square distribution $\chi^{-2}(\upsilon, S)$ was used as prior distribution for the variance of the marker effects with hyperparameters degrees of freedom υ and scale S, where $\upsilon = 4.234$ and $S = 0.0429$ (see Meuwissen et al. 2001). In this study, we used "BLR" package (de los Campos et al. 2013) to implement BayesB model and adopted the default values for υ and S.

RKHS The general form of the RKHS method is defined as

$$\mathrm{y} = \mu + \mathrm{K}_h\beta + \epsilon, \quad 5$$

where K_h is a kernel function, which can be used to map the input data to a high dimensional space where the data can be more easily separated, β and ϵ are assumed to have independent prior distributions $\beta \sim NIID\left(0,\ \mathrm{K}_h\sigma_\beta^2\right)$ and $\epsilon \sim NIID\left(0,\ I\sigma_e^2\right)$. RKHS has been used for spatial smoothing, regression, and classification, in which the reproducing kernel (RK) is one of the central elements of model specification. Here, we selected the multi-kernel function and implemented the method in the R package BGLR (de Los Campos et al. 2009b).

SNN The Single Hidden Layer Feed Forward Neural Networks for GS is introduced by Gianola et al. (2011):

$$y_i = \mu + \sum_{k=1}^{S} W_k g_k\left(b_k + \sum_{j=1}^{p} x_{ij}\beta_j^{[k]}\right) + \epsilon_i. \quad 6$$

In terms of genome-enabled prediction using [3], in the hidden layer, the genomic covariates x_{ij} (for $j = 1, \cdots, p$) of an individual i (for $i = 1, \cdots, n$) are linearly combined with a vector of input weights $\beta_j^{[k]}$ that are specified in the training phase, plus an intercept (in NN's terminology also called "bias") b_k with $k = 1, \cdots, S$ denoting a neuron. The resulting linear score is then transformed using an activation function $g_k(.)$ to produce the output of the single hidden neuron. To model non-linear relationship between phenotype and input, the tangent hyperbolic function ($\tanh(x) = \frac{2}{[1+\exp(-2x)]} - 1$) can be used in the hidden neurons. In the output layer, the S genotype-derived basis functions, resulting from the hidden layer, are also linearly combined by using the W_1, W_2, $\cdots$, W_S weights.

We obtain an estimate of sparse structure of model [6] by minimizing the negative logarithm of likelihood of the data with sparsity enforcing L_1-norm penalty on parameters $\left\{W_k, b_k, \beta_j^{[k]}\right\} (k = 1, \cdots, S;\ \ j = 1, \cdots, p)$as follows:

$$\min_{W_k, b_k, \beta_j^{[k]}} \tilde{F}\left(W_k, b_k, \beta_j^{[k]}\right) \triangleq \widehat{\mathcal{L}}\left(W_k, b_k, \beta_j^{[k]}\right) + \left(\sum_{k=1}^{S}\sum_{j=1}^{p}\lambda_{k,j}|\beta_j^{[k]}| + \sum_{k=1}^{S}\lambda_k|b_k| + \sum_{k=1}^{S}\lambda_k|W_k|\right),$$

where the approximate square error.

$\widehat{\mathcal{L}}\left(W_k, b_k, \beta_j^{[k]}\right) = \sum_{i=1}^{n}\left(\sum_{k=1}^{S} W_k g_k\left(b_k + \sum_{j=1}^{p} x_{ij}\beta_j^{[k]}\right) - y_i\right)^2$, $\lambda_{k,j} (\lambda_{k,j} > 0)$ and $\lambda_k (\lambda_k > 0)$ are Lagrange multipliers that determine the amount of sparsity in $\beta_j^{[k]}$, W_k , and b_k. For SNN, we calculated the noise-to-signal ratio $\lambda = \sigma_\beta^2/\sigma_e^2$ and implemented the SNN method in the R package snnR (Wang et al. 2018).

Predictability and Heritability

The predictability for scallop hybrid performance was evaluated using a tenfold cross-validation, where the sample was randomly partitioned into ten parts with four parts being used to estimate parameters and the remaining part being predicted. Finally, all parts were predicted once and used four times to estimate parameters. The predictability is defined as the correlation coefficient between the observed and predicted phenotypic values. The predictability may be affected by how the sample is partitioned into the tenfold. Therefore, we replicated the cross-validation analysis 100 times to achieve the average prediction results of these replicates. In order to identify the impacts of training population size and marker number on predictability, we used different subsets of training population

and markers to evaluate the predictability. Data accessibility: phenotype and sequence data are available from: “ http://mgb.ouc.edu.cn/cfbase/html/download.php .”

To estimate the variance components, we used GCTA version 1.24.2 (Yang et al. 2011) to estimate the proportion of phenotypic variance explained by the genotyped SNPs. First, GCTA was used to create the genetic relationship matrix (GRM) for estimating the pair-wise genetic relationship between individuals. Then, we estimated univariate heritabilities of complex traits of scallop by the restricted maximum likelihood method in GCTA. Meanwhile, σ_β^2 and σ_e^2 being the residual and marker variance component estimates obtained by SNP-based heritability using GCTA (Yang et al. 2011).

Funding Information

~~This study is supported by the National Natural Science Foundation of China (31772844 and U1706203) and Fundamental Research Funds for the Central Universities (201762001 and 201564009).~~We acknowledge the grant support from the National Natural Science Foundation of China (31772844, U1706203, 31630081), the Major basic research projects of Shandong Natural Science Foundation (2018A07), and the Fundamental Research Funds for the Central Universities (201762001, 201841001).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest.

References

Abdelrahman H, ElHady M, Alcivar-Warren A, Allen S, Al-Tobasei R, Bao L et al (2017) Aquaculture genomics, genetics and breeding in the United States: current status, challenges, and priorities for future research. BMC Genomics 18(1):191

Aliloo H, Pryce JE, Gonzalezrecio O, Cocks BG, Hayes B (2015) Validation of markers with non-additive effects on milk yield and fertility in Holstein and Jersey cows. BMC Genet 16:89–89

Bernardo R, Yu J (2007) Prospects for genome-wide selection for quantitative traits in maize. Crop Sci 47:1082–1090

Browning B, Browning S (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated

individuals. Am J Hum Genet 84:210–223

Crossa J, Perez P, Hickey J, Burgueno J, Ornella L, CeronRojas J et al (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112:48–60

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031

De Los Campos G, Naya H, Gianola D, Crossa J, Legarra A et al (2009a) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385

De Los Campos G, Gianola D, Rosa GJM (2009b) Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J Anim Sci 87:1883–1887

De Los Campos G, Perez P, Vazquez AI, Crossa J (2013) Genome-enabled prediction using the BLR (Bayesian linear regression) R-package. Methods Mol Biol 1019:299–320

Desta ZA, Ortiz R (2014) Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci 19:592–601

Dong L, Xiao S, Chen J, Wan L, Wang Z (2016) Genomic selection using extreme phenotypes and pre-selection of SNPs in large yellow croaker (*Larimichthys crocea*). Mar Biotechnol 18:575–583

Dou J, Li X, Fu Q, Jiao W, Li Y, Li T, Wang Y, Hu X, Wang S, Bao Z (2016) Evaluation of the 2b-rad method for genomic selection in scallop breeding. Sci Rep 6:19244

Ehret A, Hochstuhl D, Gianola D, Thaller G (2015) Application of neural networks with back-propagation to genome-enabled prediction of complex traits in Holstein-Friesian and German Fleckvieh cattle. Genet Sel Evol 47:22

Falconer D, Mackay T (1996) Introduction to quantitative genetics, 4th edn. Benjamin Cummings, Essex England

Ge J, Li Q, Yu H, Kong L (2015) Identification of single-locus PCR-based markers linked to shell background color in the Pacific Oyster (*Crassostrea*

gigas). Mar Biotechnol 17:655–662

Gelandi P, Kowalski BR (1986) Partial least-squares regression: a tutorial. Anal Chim Acta 185:1–17

Gonzalez-Recio O, Gianola D, Long N, Wiegel K, Rosa GJM et al (2008) Non parametric methods for incorporating genomic information into genetic evaluation: an application to mortality in broilers. Genetics 178:2305–2313

Goddard M, Hayes B (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10:381–391

Gianola D, Okut H, Weigel KA, Rosa GJM (2011) Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet 12:87–100

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397

Hayes B, Bowman P, Chamberlain AJ, Goddard M (2009) Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–443

Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52:146–160

Hill WG (2013) Selective breeding. Brenners Encyclopedia of Genetics 1:371–373

Howard R, Carriquiry AL, Beavis WD (2014) Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 (Bethesda, Md) 4:1027–1046

Ibarra A, Ramirez J, Ruiz C, Cruz P, Avila S (1999) Realized heritabilities and genetic correlation after dual selection for total weight and shell width in catarina scallop (*Argopecten circularis*). Aquaculture 175:227–241

Kessuwan K, Kubota S, Liu Q, Sano M, Okamoto N, Sakamoto T, Yamashita H, Nakamura Y, Ozaki A (2016) Detection of growth-related quantitative trait loci and high-resolution genetic linkage maps using simple sequence repeat markers in the kelp grouper (*Epinephelus bruneus*). Mar Biotechnol 18:57–84

Li HL, Gu XH, Li BJ, Chen CH, Lin HR, Xia JH (2017a) Genome-wide QTL analysis identified significant associations between hypoxia tolerance and mutations in the GPR132 and ABCG4 genes in Nile tilapia. Mar Biotechnol 19:441–453

Li Y, Sun X, Hu X, Xun X, Zhang J, Guo X, Jiao W, Zhang L, Liu W, Wang J, Li J, Sun Y, Miao Y, Zhang X, Cheng T, Xu G, Fu X, Wang Y, Yu X, Huang X, Lu W, Lv J, Mu C, Wang D, Li X, Xia Y, Li Y, Yang Z, Wang F, Zhang L, Xing Q, Dou H, Ning X, Dou J, Li Y, Kong D, Liu Y, Jiang Z, Li R, Wang S, Bao Z (2017b) Scallop genome reveals molecular adaptations to semi-sessile life and neurotoxins. Nat Commun 8(1):1721

Liang J, Zhang G, Zheng H (2010) Divergent selection and realized heritability for growth in the Japanese scallop, *Patinopecten yessoensis jay*. Aquac Res 41:1315–1321

Lin G, Wang L, Ngoh ST, Ji L, Orbán L, Yue GH (2018) Mapping QTL for Omega-3 content in hybrid saline tilapia. Mar Biotechnol 20:10–19

Liu P, Wang L, Wan ZY, Ye BQ, Huang S, Wong SM, Yue GH (2016) Mapping QTL for resistance against viral nervous necrosis disease in Asian seabass. Mar Biotechnol 18:107–116

Liu Y, Lu S, Liu F, Shao C, Zhou Q, Wang N et al (2018) Genomic selection using BayesCπ and GBLUP for resistance against Edwardsiella tarda in Japanese flounder (*Paralichthys olivaceus*). Mar Biotechnol. https://doi.org/10.1007/s10126-018-9839-z

Maenhout S, De Baets B, Haesaert G, Van Bockstaele E (2007) Support vector machine regression for the prediction of maize hybrid performance. Theor Appl Genet 115:1003–1013

Meuwissen THE, Hayes B, Goddard M (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Negrín-Báez D, Negrín-Báez D, Rodríguez-Ramilo ST, Afonso JM, Zamorano MJ (2016) Identification of quantitative trait loci associated with the skeletal deformity LSK complex in gilthead Seabream (*Sparus aurata* L.). Mar Biotechnol 18:98–106

Odegard J, Moen T, Santi N, Korsvoll S, Kjoglum S, Meuwissen THE (2014) Genomic prediction in an admixed population of Atlantic salmon (*salmo*

salar). Front Genet 5:402

Piepho HP (2009) Ridge regression and extensions for genome wide selection in maize. Crop Sci 49:1165–1176

Rodríguez-Ramilo S, García-Cortés T, Rodríguez de Cara M (2015) Artificial selection with traditional or genomic relationships: consequences in coancestry and genetic diversity. Front Genet 6:127

Sawayama E, Tanizawa S, Kitamura SI, Nakayama K, Ohta K, Ozaki A, Takagi M (2017) Identification of quantitative trait loci for resistance to RSIVD in Red SeaBream (*Pagrus major*). Mar Biotechnol 19:601–613

Sawayama E, Noguchi D, Nakayama K, Takagi M (2018) Identification, characterization, and mapping of a novel SNP associated with body color transparency in Juvenile Red Sea Bream (*Pagrus major*). Mar Biotechnol 20:481–489

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B 58:267–288

Tsai HY, Hamilton A, Tinch AE, Guy DR, Gharbi K, Stear MJ, Matika O, Bishop SC, Houston RD (2015) Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. BMC Genomics 16:969

Vallejo RL, Leeds TD, Fragomeni BO, Gao G, Hernandez AG, Misztal I, Welch TJ, Wiens GD, Palti Y (2016) Evaluation of genome-enabled selection for bacterial cold water disease resistance using progeny performance data in rainbow trout: insights on genotyping methods and genomic prediction models. Front Genet 7:96

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Vitezica ZG, Varona L, Legarra A (2013) On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195(4):1223–1230

Wang L, Bai B, Huang S, Liu P, Wan ZY, Ye B, Wu J, Yue GH (2017a) QTL mapping for resistance to Iridovirus in Asian seabass using genotyping-by-sequencing. Mar Biotechnol 19:517–527

Wang L, Liu P, Huang S, Ye B, Chua E, Wan ZY, Yue GH (2017b) Genome-wide association study identifies loci associated with resistance to viral nervous necrosis disease in Asian seabass. Mar Biotechnol 19:255–265

Wang S, Meyer E, Mckay JK, Matz MV (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. Nat Methods 9:808–810

Wang S, Zhang J, Jiao W, Li J, Xun X, Sun Y, Guo X, Huan P, Dong B, Zhang L, Hu X, Sun X, Wang J, Zhao C, Wang Y, Wang D, Huang X, Wang R, Lv J, Li Y, Zhang Z, Liu B, Lu W, Hui Y, Liang J, Zhou Z, Hou R, Li X, Liu Y, Li H, Ning X, Lin Y, Zhao L, Xing Q, Dou J, Li Y, Mao J, Guo H, Dou H, Li T, Mu C, Jiang W, Fu Q, Fu X, Miao Y, Liu J, Yu Q, Li R, Liao H, Li X, Kong Y, Jiang Z, Chourrout D, Li R, Bao Z (2017c) Scallop genome provides insights into evolution of bilaterian karyotype and development. Nat Ecol Evol 1(5):120

Wang Y, Mi X, Rosa GJM, Chen Z, Lin P, Wang S, Bao Z (2018) Technical note: an R package for fitting sparse neural networks with application in animal breeding. J Anim Sci 96:2016–2026

Yang J, Lee S, Goddard M, Visscher P (2011) GCTA: a tool for genome wide complex trait analysis. Am J Hum Genet 88:76–82

Zhao Y, Peng W, Guo H, Chen B, Zhou Z, Xu J, Zhang D, Xu P (2018) Population genomics reveals genetic divergence and adaptive differentiation of Chinese Sea Bass (*Lateolabrax maculatus*). Mar Biotechnol 20:45–59

Zheng H, Zhang G, Liu X, Guo X (2006) Sustained response to selectionin an introduced population of the hermaphroditic bay scallop *argopecten irradians irradians lamarck* (1819). Aquaculture 255:579–585

Zhong S, Dekkers JCM, Fernando RL, Jannink JL (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182:355–364

Zhong X, Wang X, Zhou T, Jin Y, Tan S, Jiang C, Geng X, Li N, Shi H, Zeng Q, Yang Y, Yuan Z, Bao L, Liu S, Tian C, Peatman E, Li Q, Liu Z (2017) Genome-wide association study reveals multiple novel QTL associated with low-oxygen tolerance in hybrid catfish. Mar Biotechnol 19(4):379–390

Jannink J L, Lorenz A J, Iwata H. (2010) Genomic selection in plant breeding: from theory to practice. Briefings in functional genomics 9(2): 166-177

Li Y, Wang R, Xun X, Wang J, Bao L, Thimmappa R, Ding J, Jiang J, Zhang L, Li T, Lv J, Mu C, Hu X, Zhang L, Liu J, Li Y, Yao L, Jiao W, Wang Y, Lian S, Zhao Z, Zhan Y, Huang X, Liao H, Wang J, Sun H, Mi X, Xia Y, Xing Q, Lu W, Osbourn A, Zhou Z, Chang Y, Bao Z, Wang S, (2018) Sea cucumber genome provides insights into saponin biosynthesis and aestivation regulation. Cell Discovery 4 (1)

Wang S, Liu P, Lv J, Li Y, Cheng T, Zhang L, Xia Y, Sun H, Hu X, Bao Z, (2016) Serial sequencing of isolength RAD tags for cost-efficient genome-wide profiling of genetic and epigenetic variations. Nature Protocols 11 (11):2189-2200

Wang S, Lv J, Dou J, Lu Q, Zhang L & Bao Z. (2017d) Chapter 19. Genotyping by sequencing and data analysis: RAD and 2b-RAD sequencing. In: Liu Z (ed). Bioinformatics in Aquaculture. NJ: John Wiley & Sons Publishing Ltd. Pp. 338-355.

Jiao W, Fu X, Dou J, Li H, Su H, Mao J, Yu Q, Zhang L, Hu X, Huang X, Wang Y, Wang S, Bao Z, (2014) High-Resolution Linkage and Quantitative Trait Locus Mapping Aided by Genome Survey Sequencing: Building Up An Integrative Genomic Framework for a Bivalve Mollusc. DNA Research 21 (1):85-101